Correlation analysis

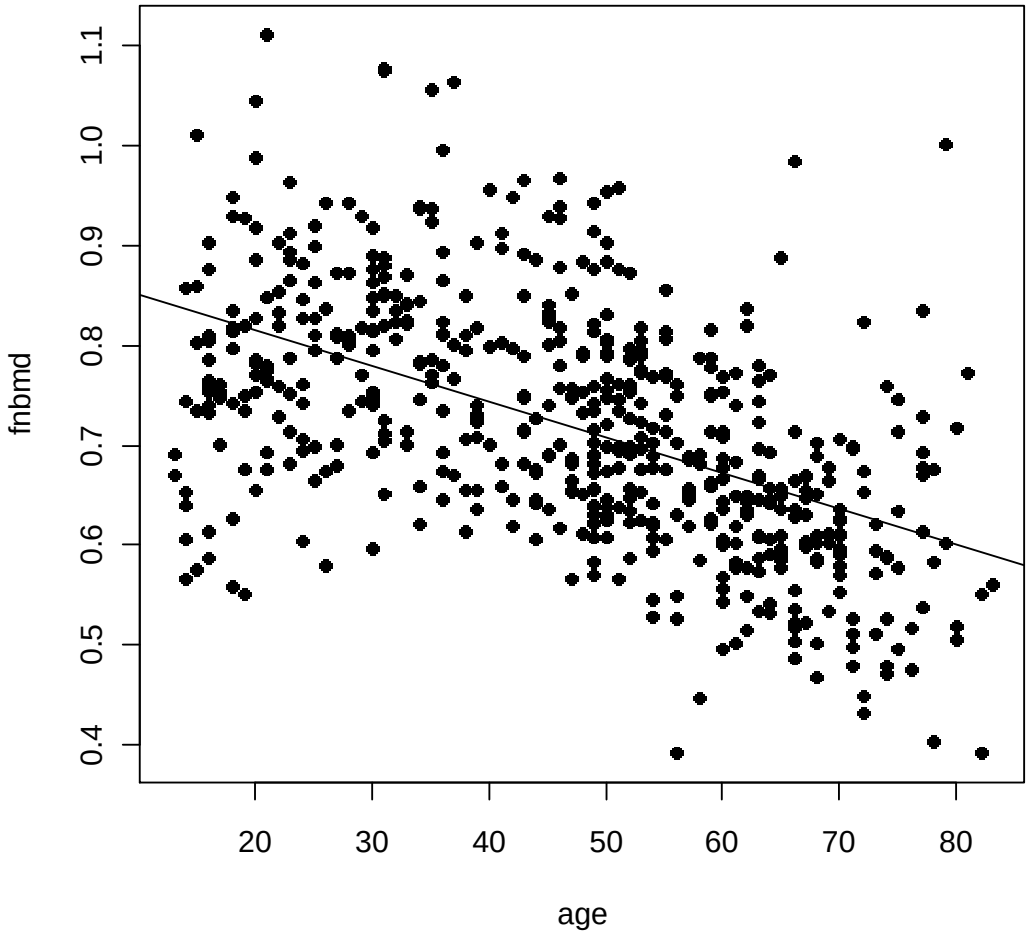
Tuan V. Nguyen
Professor and NHMRC Senior Research Fellow
Garvan Institute of Medical Research
University of New South Wales
Sydney, Australia

What we are going to learn …

- Examples (BMD and age, BMD and weight)
- Formulating question
- Covariance and correlation
- R analysis
- Interpretation
- Linear regression

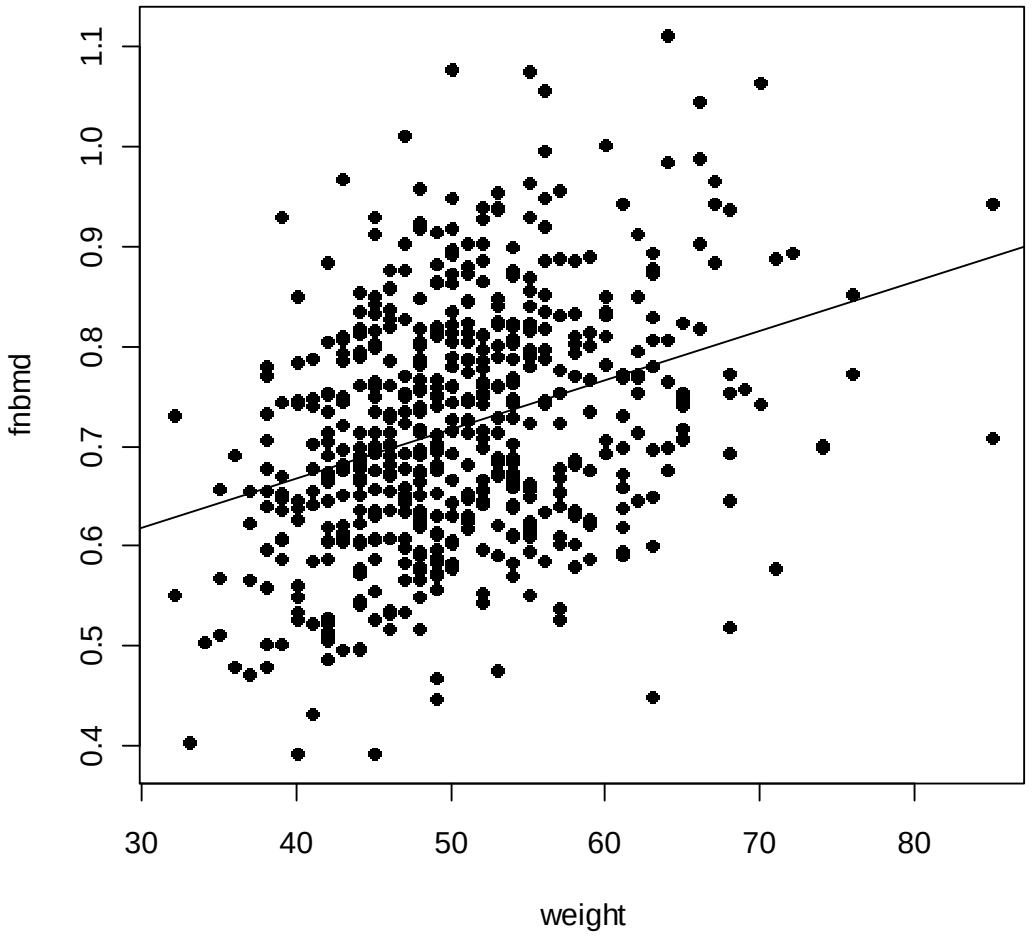
Femoral neck bone density and age

```
women = subset(vd, sex==2)
plot(fnbmd ~ age, pch=16)
abline(lm(fnbmd ~ age))
```



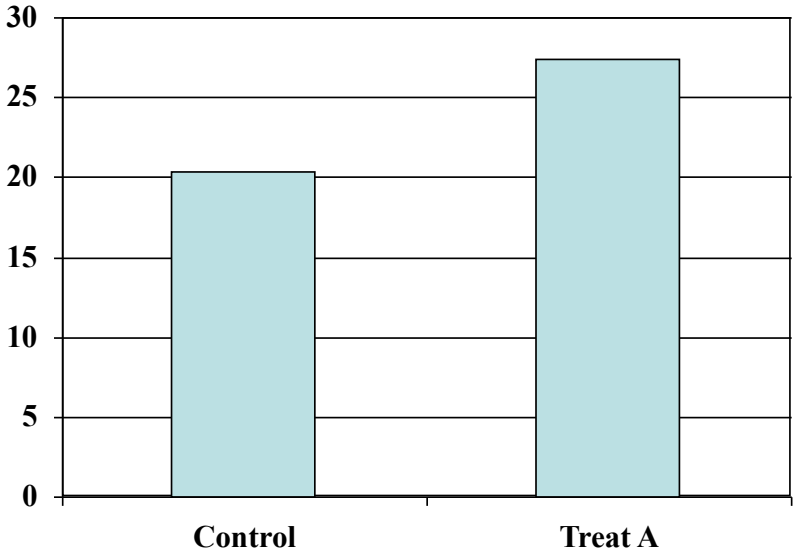
Weight and femoral neck bone density

```
plot(fnbmd ~ weight, pch=16)
abline(lm(fnbmd ~ weight))
```

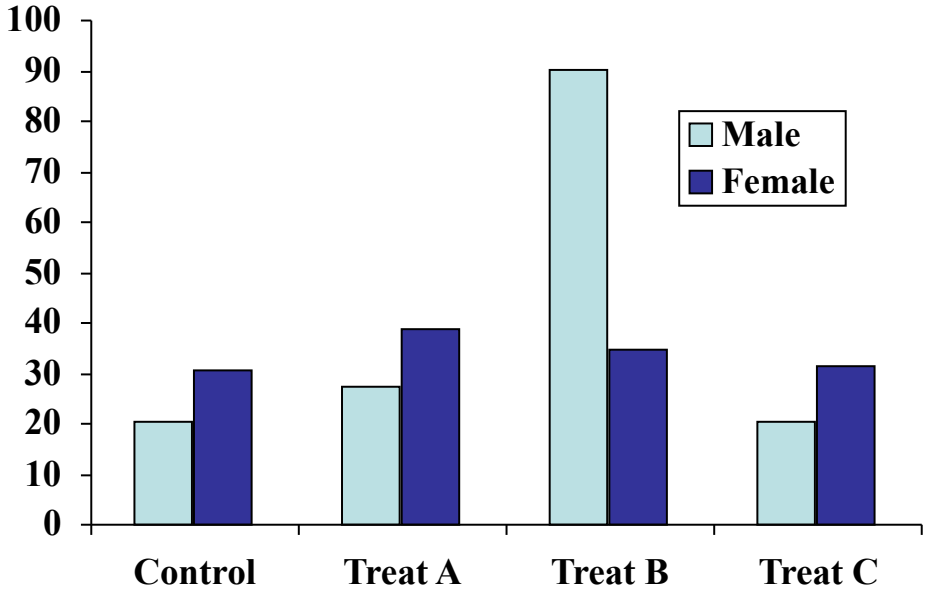


When to consider correlation?

t-test



or ANOVA

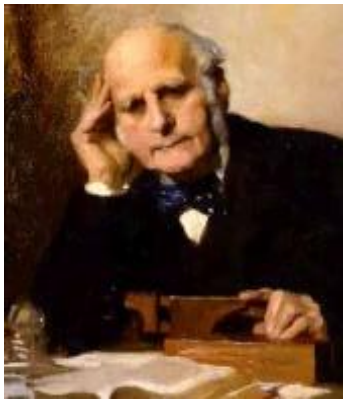


When the independent variable is *categorical*

When to consider correlation?

- *Relationship* between 2 continuous variables
- Degree of co-variation
- Predition ?

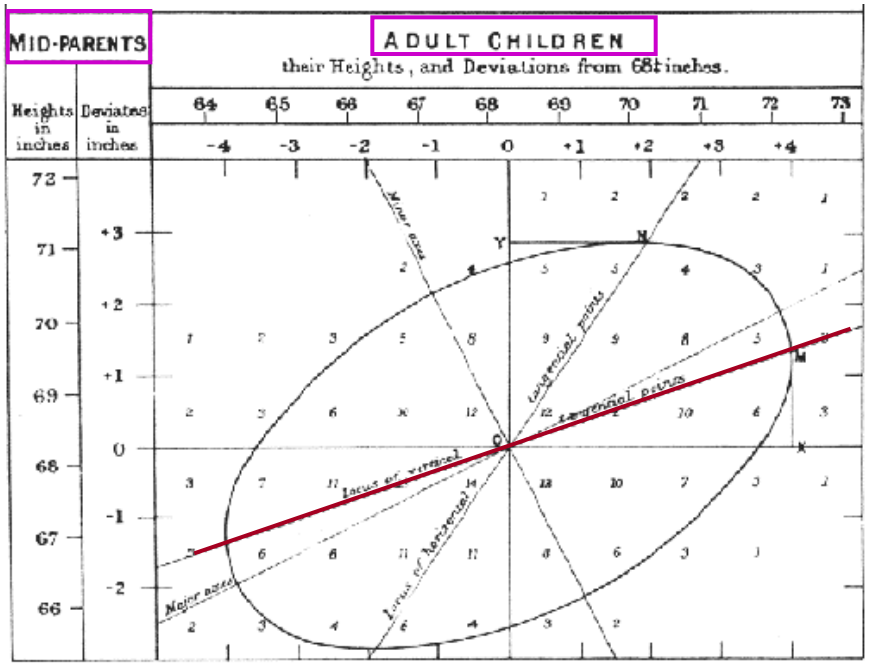
Meet Sir Francis Galton (16/2/1822 – 17/1/1911)



Research interest:

"Those qualifications of intellect and disposition which … lead to reputation"

Didn't have data on "intelligence" so instead studied HEIGHT



Galton's conclusions:

- **Nature dominates:** "families of reputation were much more likely than ordinary families to produce offspring of ability"
- **Recommended "judicious marriages during several generations" to "produce a highly gifted race of men"**
- His "genetic utopia": **"Bright, healthy individuals were treated and paid well, and encouraged to have plenty of children. Social undesirables were treated with reasonable kindness so long as they worked hard and stayed celibate."**

- Although a self-proclaimed genius, who wrote that he could read @2½, write/do arithmetic @4, and was comfortable with Latin texts @8, **he couldn't figure out how to model these data(!)**
- He went to JD Dickson, a mathematician at Cambridge, who formalized the relationship by developing what we now know as **linear regression**

How do we describe the linear relationship?

- Let *X* and *Y* be two random variables from a sample of *n* obervations.
- Measure of variability of *x* and *y:* **variance**

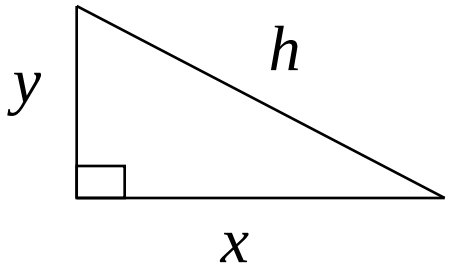
$$\text{var}(x) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \qquad \text{var}(y) = \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n-1}$$

- We need a measure of **covariation (covariance)** between *X* and *Y.*
- Covariance is the average of **product** of *X* and *Y*

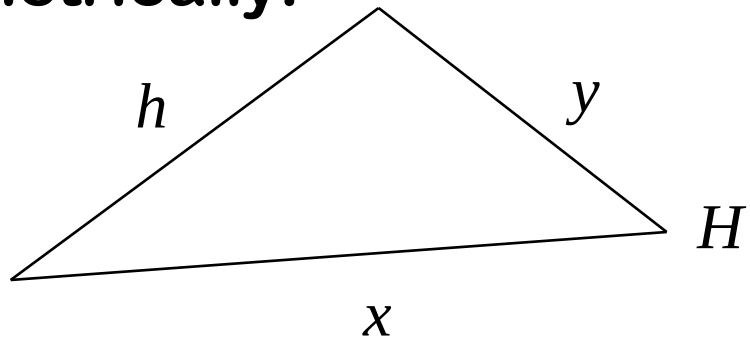
$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Variance and covariance: geometry

- **The independence or dependence between *x* and *y* can be represented geometrically:**



$$h^2 = x^2 + y^2$$



$$h^2 = x^2 + y^2 - 2xy\cos(H)$$

Covariance

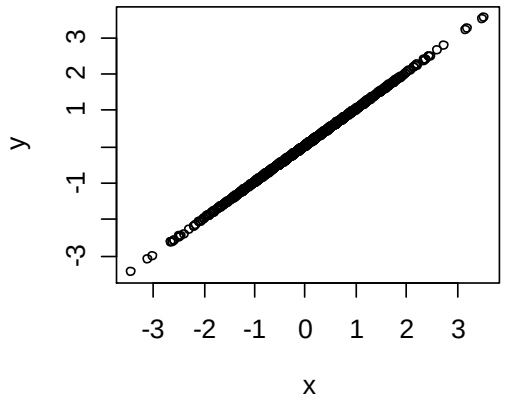
Meaning of variance and covariance

- Variance is always positive
- If covariance = 0, *X* and *Y* are independent
- Covariance is sum of cross-products: can be positive or negative

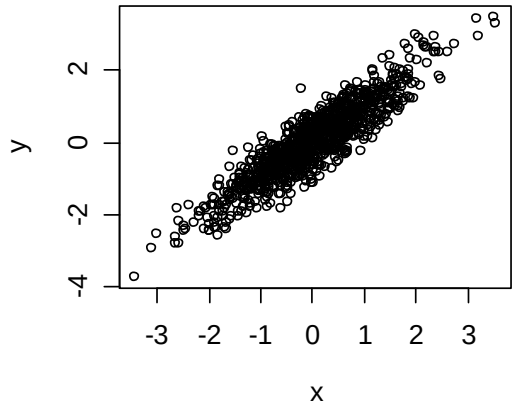
- Negative covariance = deviations in the two distributions in are opposite directions
- Positive covariance = deviations in the two distributions in are in the same direction
- **Covariance = a measure of strength of association**

Some correlations

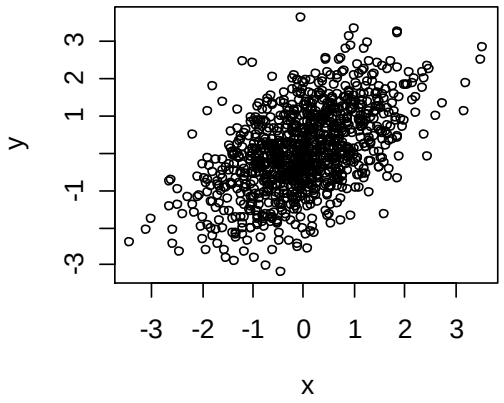
r = 0.99



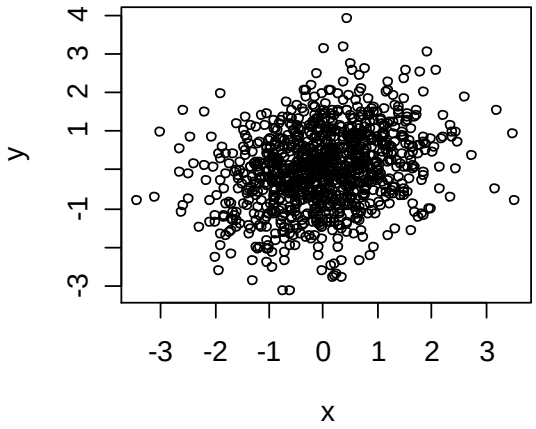
r = 0.90



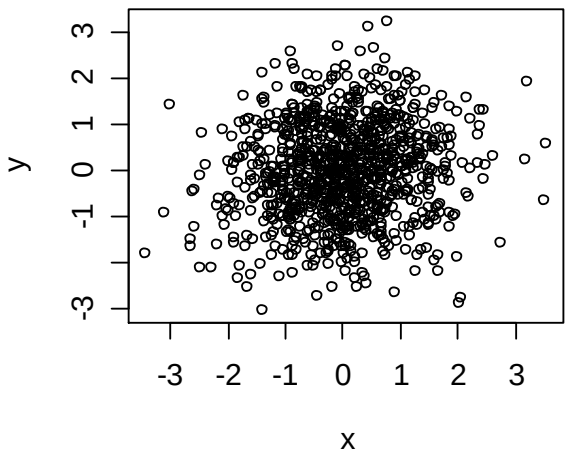
r = 0.50



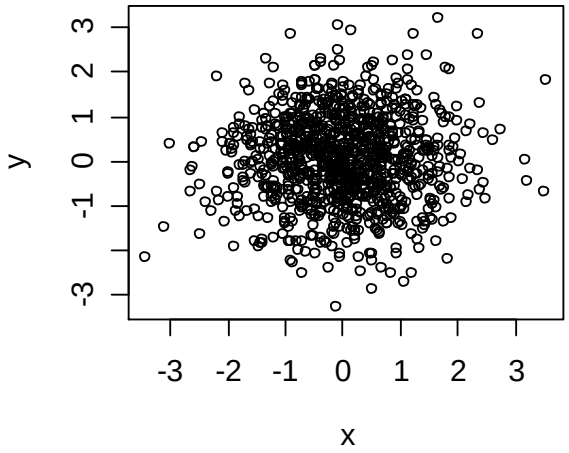
r = 0.25



r = 0.10



r = 0.01



Our aim of analysis

- To estimate the *coefficient of correlation (r)*
- To test whether $r = 0$?

Estimate of correlation coefficient *r*

- Covariance is unit-depenent.
- **Coefficient of correlation** (*r*) between *X* and *Y* is a **standardized covariance**
- *r* is defined by:

$$r = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{var}(x) \times \mathrm{var}(y)}} = \frac{\mathrm{cov}(x, y)}{SD_x \times SD_y}$$

Test of hypothesis of correlation

- Hypothesis: H_o: $r = 0$ versus H_o: r not equal to 0.
- Fisher's z-transformation: transform $r \rightarrow z$

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

- Calculate standard error of z

$$SE(z) = \frac{1}{\sqrt{n-3}}$$

- T-test:

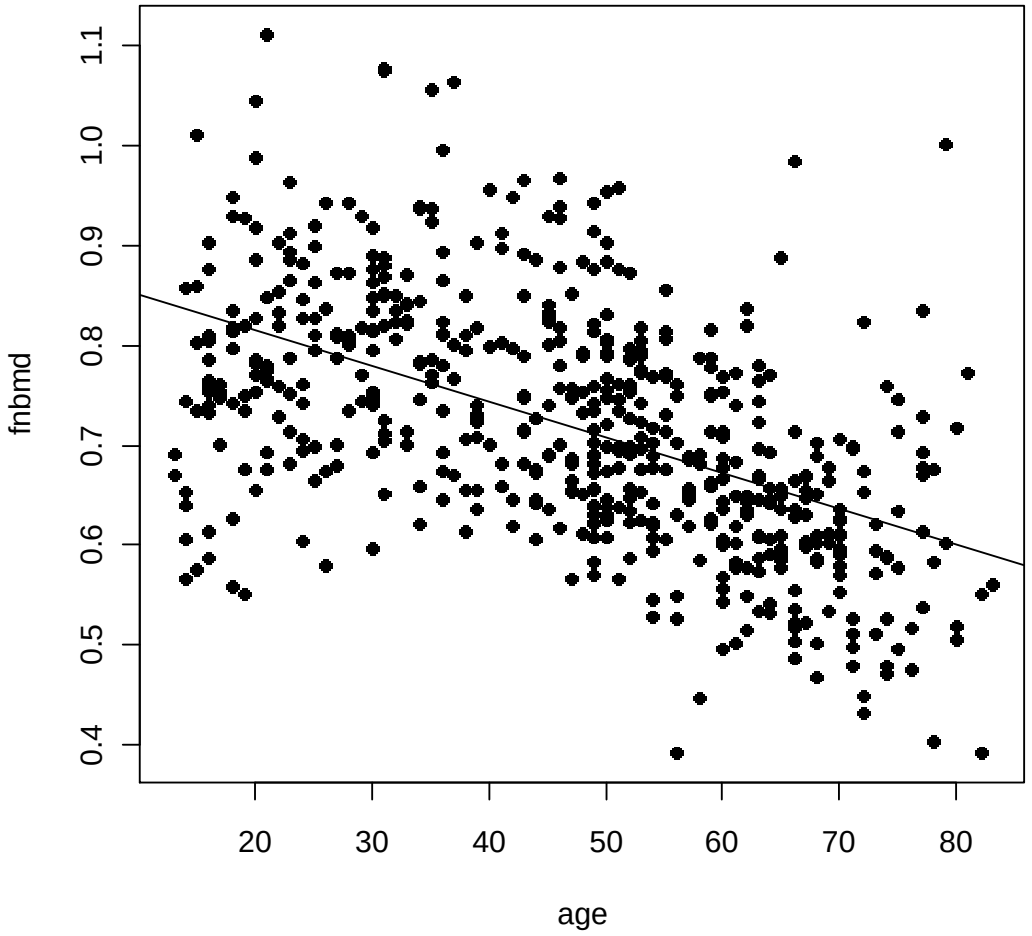
$$t = \frac{z}{SE(z)}$$

Example: bone density, age, and weight

- Cross-sectional study
- Bone mineral density (BMD) was measured at the femoral neck
- Age, weight
- Questions
  - Was there a correlation between age and BMD, weight and BMD?
  - Was the correlation statistically significant?

Femoral neck bone density and age

```
women = subset(vd, sex==2)
plot(fnbmd ~ age, pch=16)
abline(lm(fnbmd ~ age))
```



R analysis

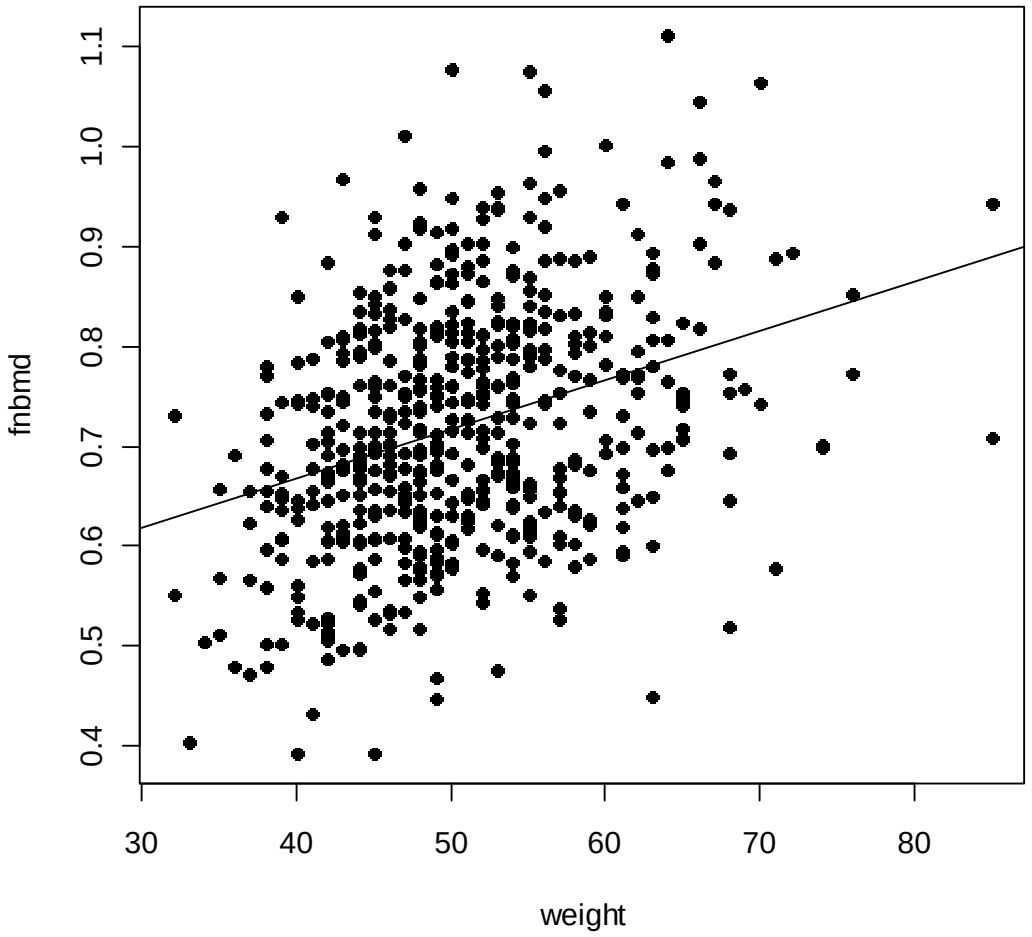
```
> cor.test(fnbmd, age)

        Pearson's product-moment correlation

data:  fnbmd and age
t = -14.4162, df = 556, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.5795310 -0.4584638
sample estimates:
       cor
-0.5216183
```

Age and BMD: an interpretation

- The coefficient of correlation between age and BMD was -0.52
- Older individuals had *lower* BMD than younger individuals
- The correlation was statistically significant ($P < 0.0001$)

Weight and femoral neck bone density

```
plot(fnbmd ~ weight, pch=16)
abline(lm(fnbmd ~ weight))
```



R analysis

```
> cor.test(fnbmd, weight)

                Pearson's product-moment correlation

data:  fnbmd and weight
t = 8.1795, df = 556, p-value = 1.998e-15
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2515705 0.3998585
sample estimates:
      cor
0.3277315
```

Weight and BMD: an interpretation

- The coefficient of correlation between weight and BMD was 0.33
- The correlation was statistically significant ($P < 0.0001$)

r and R^2

- r is the correlation coefficient
- R^2 is the coefficient of determination:

  The amount of variation of one variable can be "explained" by another variable

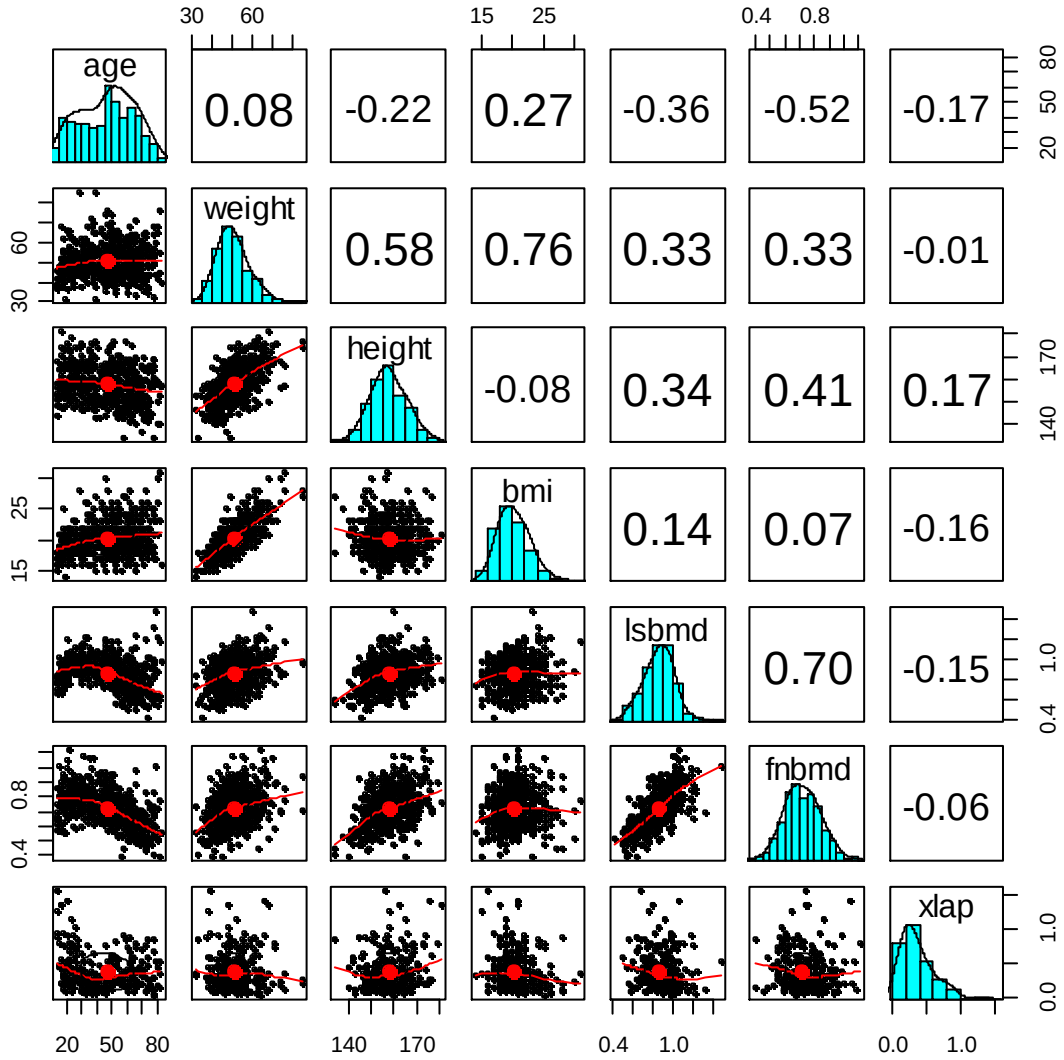
- r(weight, BMD) = 0.33 means that $R^2 = (0.33)^2 = 0.11$.

  This means that weight could explain 11% of the variation in BMD

Multi-variable correlations

- Inter-correlations among variables can be computed
- Bivariate correlation

```
library(psych)
temp = cbind(age, weight, height, bmi, lsbmd, fnbmd, xlap)
pairs.panels(temp)
```



```
> corr.test(temp)

Call:corr.test(x = temp)
```

Correlation matrix

	age	weight	height	bmi	lsbmd	fnbmd	xlap
age	1.00	0.08	-0.22	0.27	-0.36	-0.52	-0.17
weight	0.08	1.00	0.58	0.76	0.33	0.33	-0.01
height	-0.22	0.58	1.00	-0.08	0.34	0.41	0.17
bmi	0.27	0.76	-0.08	1.00	0.14	0.07	-0.16
lsbmd	-0.36	0.33	0.34	0.14	1.00	0.70	-0.15
fnbmd	-0.52	0.33	0.41	0.07	0.70	1.00	-0.06
xlap	-0.17	-0.01	0.17	-0.16	-0.15	-0.06	1.00

Probability value

	age	weight	height	bmi	lsbmd	fnbmd	xlap
age	0.00	0.07	0.00	0.00	0.00	0.00	0.01
weight	0.07	0.00	0.00	0.00	0.00	0.00	0.88
height	0.00	0.00	0.00	0.07	0.00	0.00	0.01
bmi	0.00	0.00	0.07	0.00	0.00	0.10	0.01
lsbmd	0.00	0.00	0.00	0.00	0.00	0.00	0.02
fnbmd	0.00	0.00	0.00	0.10	0.00	0.00	0.38
xlap	0.01	0.88	0.01	0.01	0.02	0.38	0.00

Be careful !

- Study on 44 university students
- Measure body mass index (BMI)
- Sexual attractiveness (SA) score

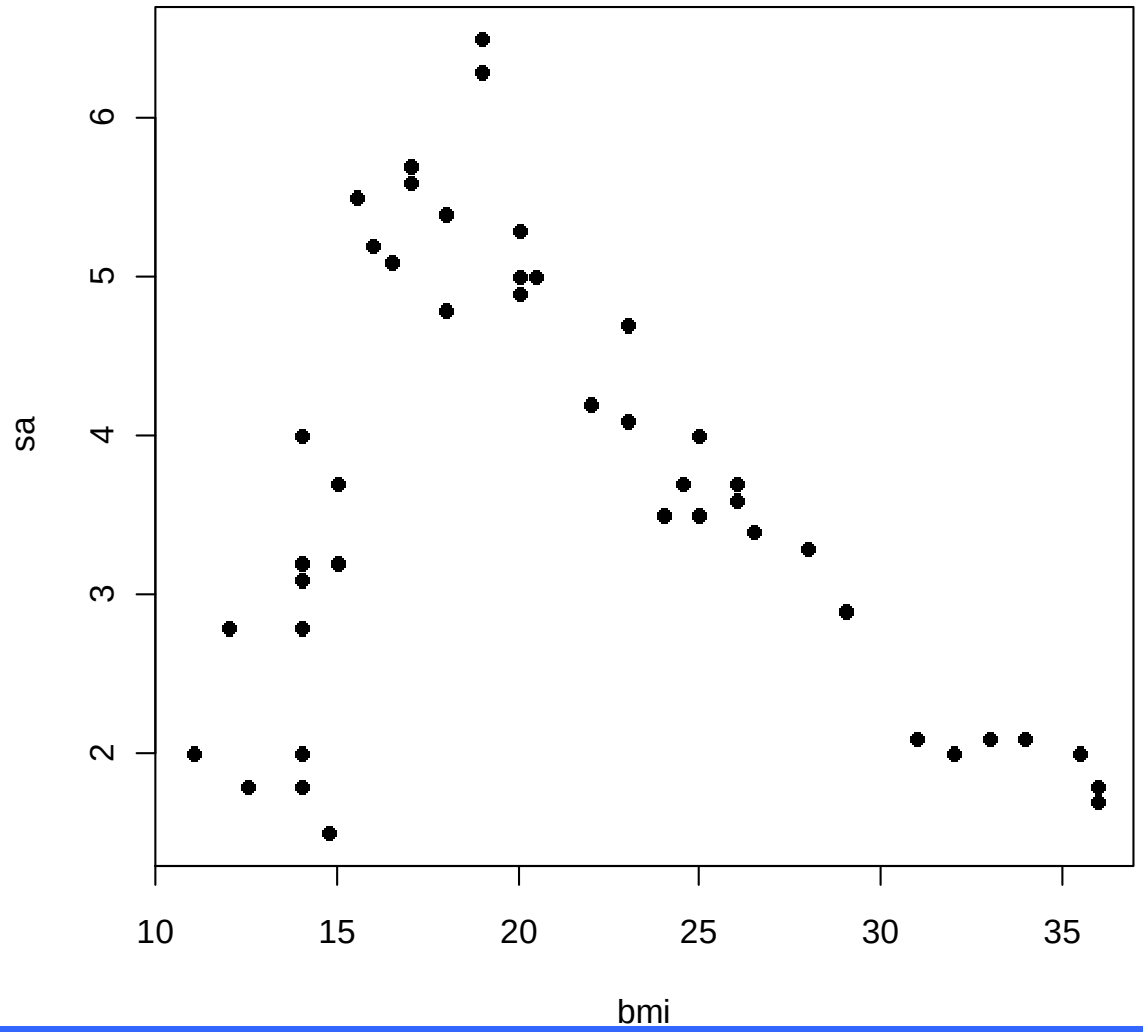
```
id <- seq(1:44)
bmi <- c(11.00, 12.00, 12.50, 14.00, 14.00, 14.00, 14.00,
         14.00, 14.00, 14.80, 15.00, 15.00, 15.50, 16.00,
         16.50, 17.00, 17.00, 18.00, 18.00, 19.00, 19.00,
         20.00, 20.00, 20.00, 20.50, 22.00, 23.00, 23.00,
         24.00, 24.50, 25.00, 25.00, 26.00, 26.00, 26.50,
         28.00, 29.00, 31.00, 32.00, 33.00, 34.00, 35.50,
         36.00, 36.00)
sa <- c(2.0, 2.8, 1.8, 1.8, 2.0, 2.8, 3.2, 3.1, 4.0, 1.5,
        3.2, 3.7, 5.5, 5.2, 5.1, 5.7, 5.6, 4.8, 5.4, 6.3,
        6.5, 4.9, 5.0, 5.3, 5.0, 4.2, 4.1, 4.7, 3.5, 3.7,
        3.5, 4.0, 3.7, 3.6, 3.4, 3.3, 2.9, 2.1, 2.0, 2.1,
        2.1, 2.0, 1.8, 1.7)
```

Negative correlation

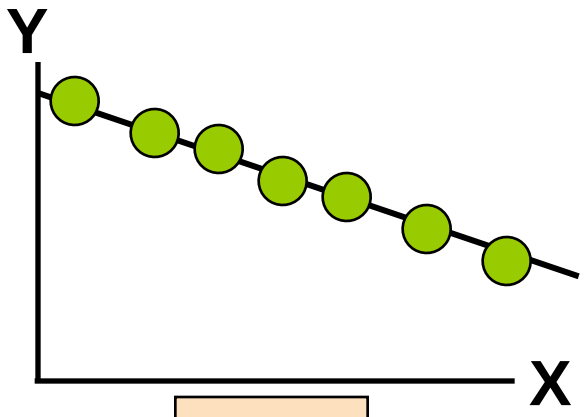
```
> cor.test(bmi, sa)

        Pearson's product-moment correlation
data:  bmi and sa
t = -2.0845, df = 42, p-value = 0.04323
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.55282250 -0.01024671
sample estimates:
       cor
-0.3061956
```

But …

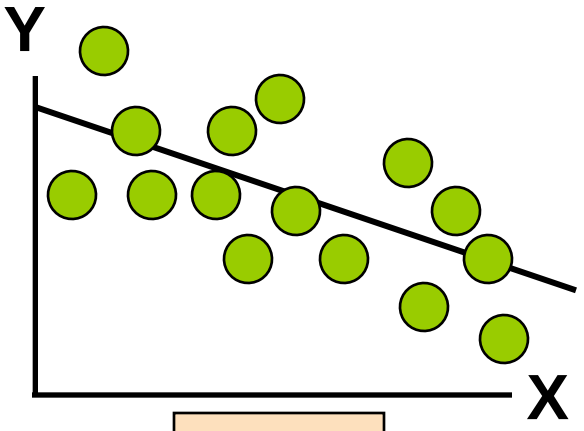
- plot(sa, bmi)



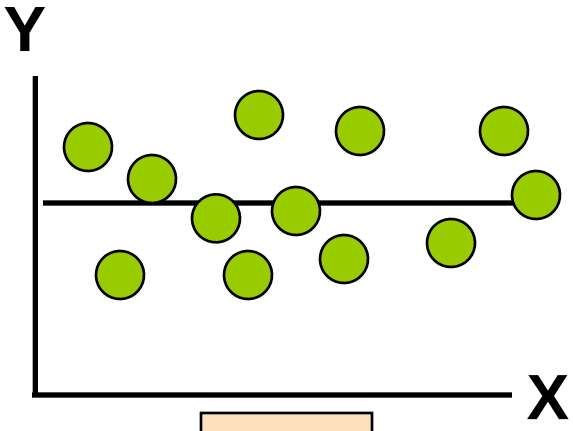
Scatter plots of data with various correlation coefficients



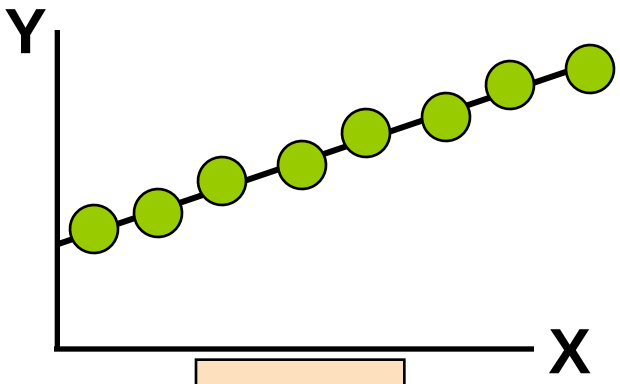
r = -1



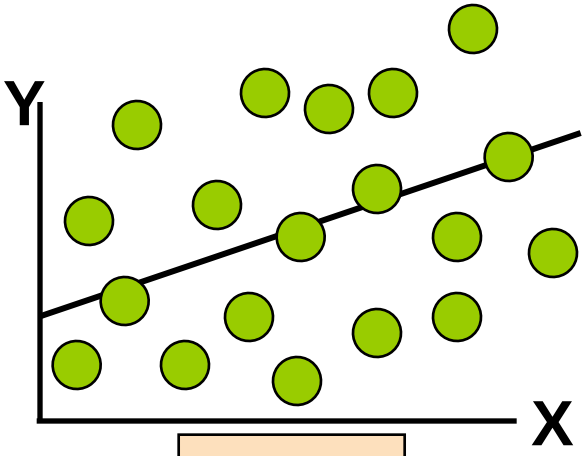
r = -.6



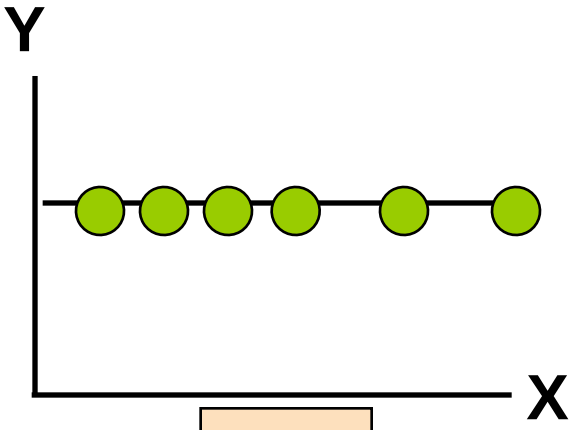
r = 0



r = +1



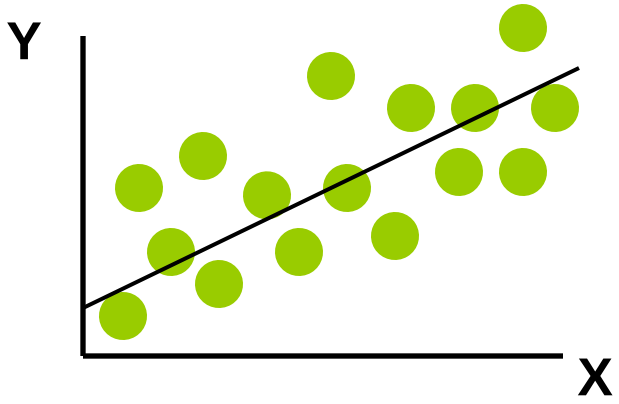
r = +.3

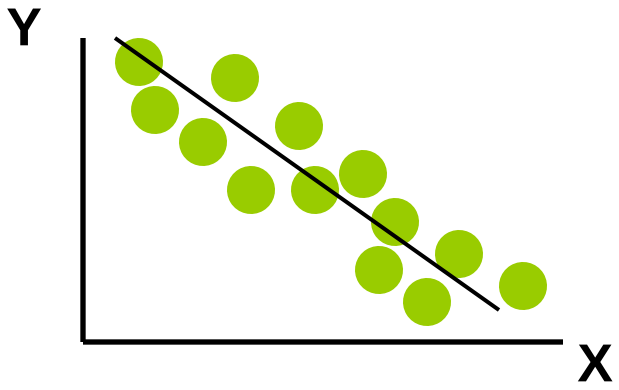


r = 0

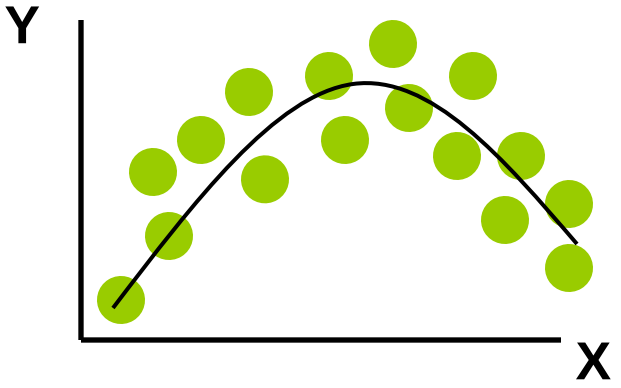
Linear correlation

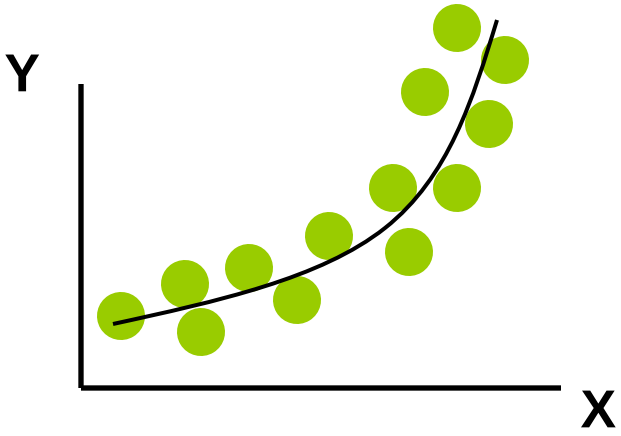
Linear relationships





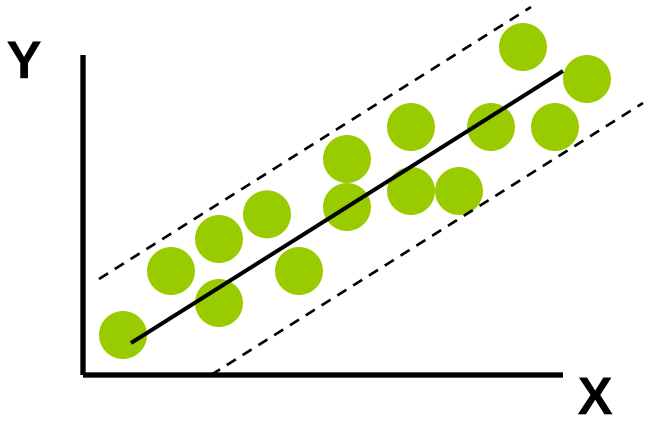
Curvilinear relationships

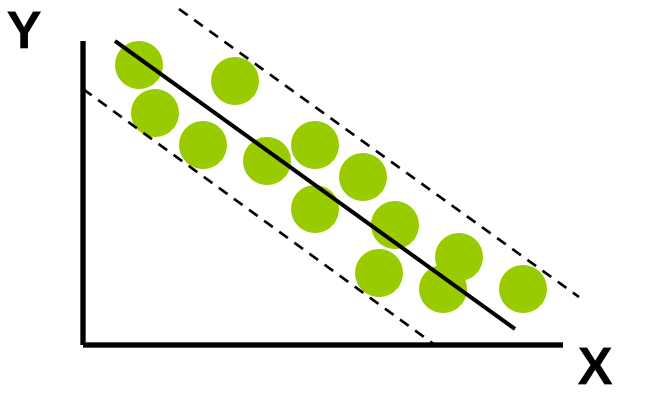




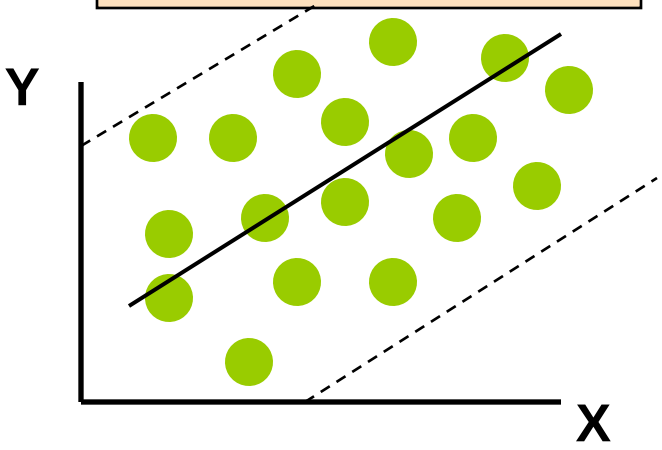
Linear correlation

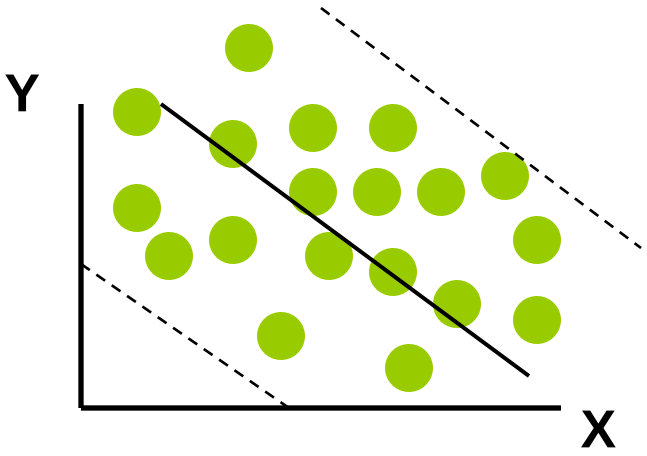
Strong relationships





Weak relationships





Summary

- A measure of strength of association between 2 continuous variables
- Normally distributed data
- Only applicable to a LINEAR relationship
- Correlation is not necessarily a measure of cause-and-effect relationship